2-O

УДК 004.652.5, 004.823

А.И. Ольшевский, А.А. Кондратьева
Государственный университет информатики и искусственного интеллекта,
г. Донецк, Украина

Описание способов представления web-сайтов в виде фреймовой модели для реализации функциональных операций в Интернет-клиентских системах

Рассмотрены вопросы структурного построения web-сайтов. Для реализации алгоритма формирования графа структуры сайта были выделены элементы, предложены модели, позволяющие сохранять информацию в реляционной базе данных для дальнейшей групповой обработки в Интернет-клиентских программах.

Введение

Одной из основных задач систем обработки данных в сети Интернет является задача поиска. Виртуальный информационный массив обладает высокой степенью динамики: каждую секунду в нем появляются новые материалы, какая-то их часть по разным причинам удаляется с серверов, другая же меняет адресацию. Это постоянное обновление с одновременным ростом объема информационного массива делает крайне сложным учет большинства документов, существующих в Интернете [1].

Важность проблемы информационного поиска в Интернете породила целую отрасль специальных поисковых инструментов. Условно их можно разделить на поисковые средства справочного типа, или просто справочники (directories), и поисковые системы в чистом виде (search engines – например, Google, Rambler и пр.). Данные системы позволяют производить поиск в определенной области сети по ключевым словам, однако не дают развёрнутых возможностей оперирования с сайтами. Поэтому актуальной является задача структурного представления web-сайтов для целей реализации над ними различных операций: сравнения двух сайтов, поиска вхождений одного сайта в другой, поиска структурных элементов одного сайта в другом и прочих. Решение этой задачи позволило бы программисту динамически формировать объект, представляющий структуру сайта, и выполнять с ним любые необходимые операции [2].

В данной статье рассматриваются основные принципы построения web-сайтов и структурное построение web-страниц. Ставится задача разработки способов представления сайта для последующего оперирования с ним как с объектом определенного типа и введения перечня возможных операций работы с таким объектом. Для решения этой задачи рассматриваются основные принципы структурного построения сайтов и некоторые возможные модели эффективного представления их структуры.

В рамках данной задачи предлагается формализация информации о структуре сайта в виде ориентированного графа и разрабатывается алгоритм формирования искомого графа по html-страницам, размещенным на сервере. Для сохранения построенного графа в реляционной базе данных приводится представление формируемой информации в виде фреймовой модели.

1 Описание операций над web-сайтами

Web-сайт состоит из ряда страниц. Таким образом, минимальной единицей web-данных является web-страница [3].

Любая программа, предназначенная для навигации по Интернет-пространству, дает возможность просматривать web-страницы, а также осуществлять поиск по ним. Однако в последнее время ведется речь не столько о web-страницах, сколько о web-сайтах. Web-сайт постепенно становится самостоятельным понятием. Такая тенденция дает возможность выделить сайт как отдельный объект данных и иметь возможность осуществлять работу не с отдельными его страницами, а со всем сайтом [3]. Это позволит производить операции с сайтами, такие, как:

– сравнение сайтов (по определенным критериям);
– сортировка сайтов (по определенным признакам);
– оценка сложности сайта;
– расширенный поиск по сайту различной информации, такой, как текст, картинки, гиперссылки, элементы управления, скрипты и др.;
– получение статистики с сайта (об определенных свойствах);
– классификация сайтов (для того, чтобы иметь возможность классифицировать сайт, необходимо ввести ряд классов).

Для представления сайта в виде объекта необходимо выделить основные элементы его структуры.

2 Формализация структуры web-сайта

Каждый сайт представляет собой набор страниц. При этом у каждого сайта имеется главная страница, с которой ведется дальнейшая навигация по сайту.

Web-страница – это html-документ, который содержит следующую информацию:

– название страницы;
– стиль оформления страницы (цвет фона, цвет и размер текста, количество фреймов на странице, размер каждого фрейма и т.д.);
– элементы управления (кнопки, переключатели, поля ввода информации и др.);
– рисунки;
– анимацию;
– информацию об используемых скриптах;
– гиперссылки (ссылки, содержащие полный путь к другой странице – по ним можно осуществлять переход на другие ресурсы).

Таким образом, практически каждый сайт организован как набор страниц, переход по которым производится с помощью ссылок [4], [5]. Исходя из подобной структуры, можно выделить несколько возможных моделей представления сайта:

– представления в виде списка;
– представление в виде графа.

Описывая сайт списком, необходимо в виде списка отобразить всю информацию о сайте: список web-страниц с перечнем их содержимого. Если представить сайт в виде графа, то вершинами графа будут являться web-страницы, а дугами – гиперссылки. Такая модель представления сайта логична и наилучшим образом отражает его структурное представление. Кроме того, она удобна для осуществления с ней операций.

Общая структура web-сайта представлена на рис. 1.

Структуру, изображенную на рис. 1, можно представить в виде ориентированного графа. Граф, представляющий структуру web-сайта, приведен на рис. 2.

Граф структуры сайта состоит из множества вершин $A = \{A_i\}$, $i = 1, 2 \ldots n$, причём вершина A_i представляет web-страницу сайта, а n – количество страниц. Главной страницей сайта всегда является A_1.

$A_i \rightarrow E_i$, т.е. каждой вершине A_i поставлено в соответствие множество исходящих дуг $E_i = \{a_{ij}\}$, $i = 1, 2 \ldots n$; $j \in \{1, 2 \ldots n\}$. Дуга a_{ij} связывает вершину A_i и A_j и соответствует гиперссылке на web-страницу A_j, находящуюся на web-странице A_j.

Таким образом, схема структуры сайта – это ориентированный граф, вершинами которого являются страницы A_i, а дугами – гиперссылки a_{ij}. Граф структуры сайта является двунаправленным, т.е. графом, в котором любые две вершины могут взаимно ссылаться друг на друга.

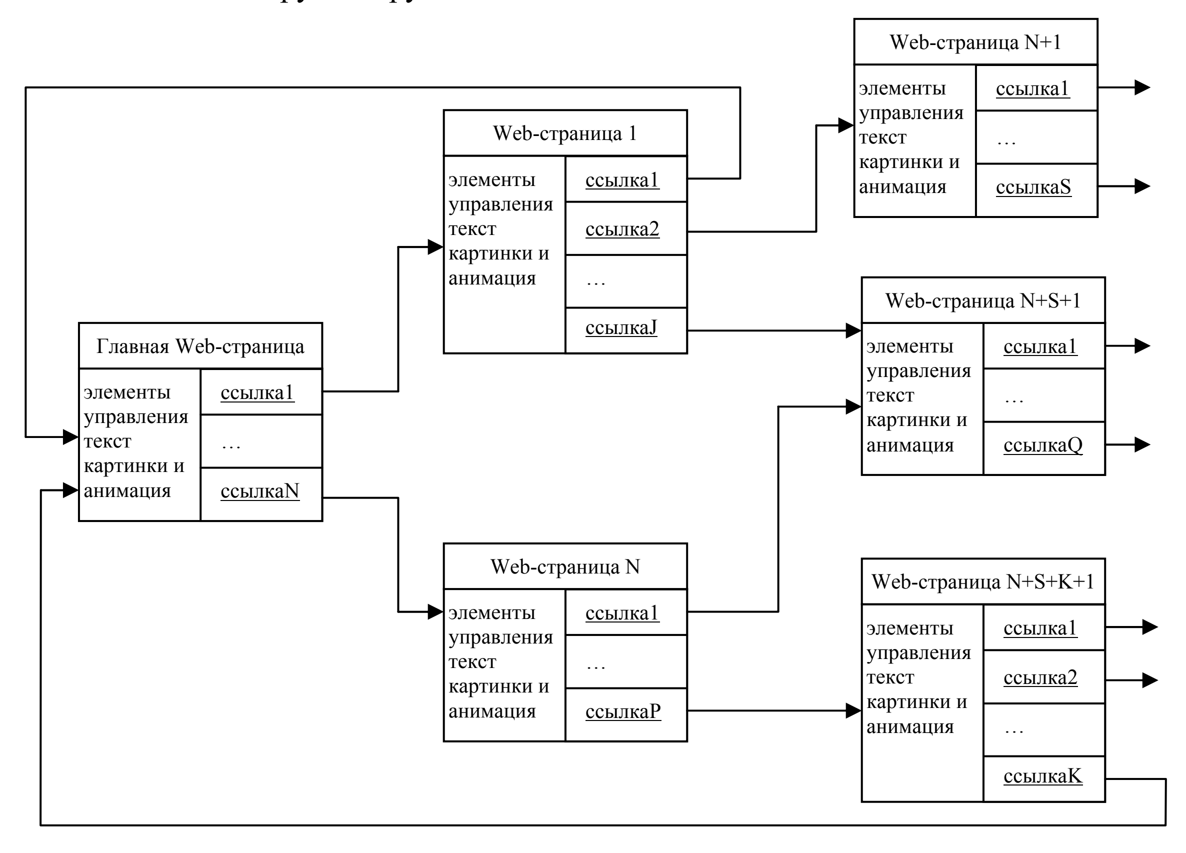


Рисунок 1 – Общая схема структуры web-сайта

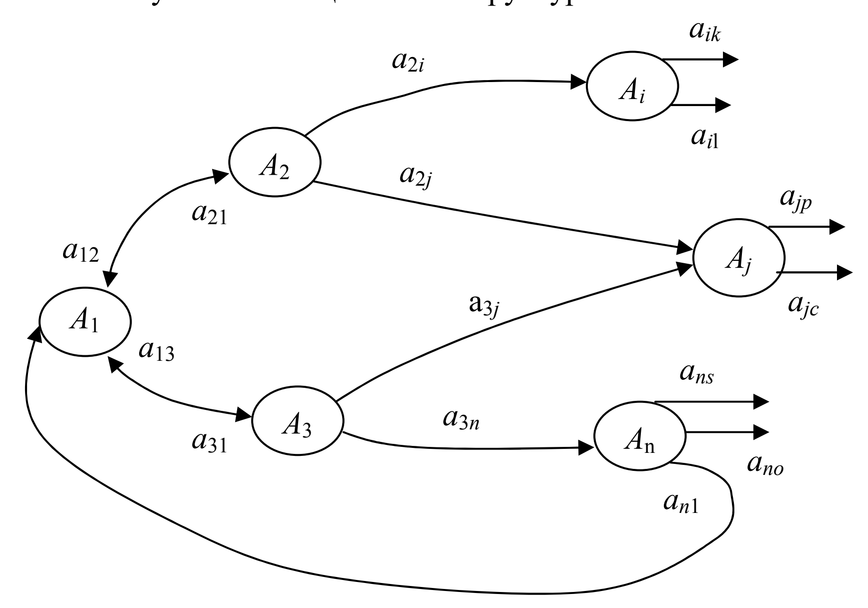


Рисунок 2 – Граф, представляющий структуру сайта

Возможны случаи, когда модель будет вырождаться в однонаправленный граф, то есть такой граф, никакие из двух вершин которого не имеют одновременно ссылки друг на друга (то есть $\forall i, j$ не существуют одновременно a_{ij} и a_{ji}).

Также возможны ситуации, когда модель сайта в виде графа не будет иметь циклов, т.е. будет вырождаться в дерево. Граф структуры сайта может быть представлен математически в виде матрицы смежности, списка или любым другим способом.

Решаемая задача сводится к восстановлению множества $\{A_i\}$ и $\{E_i\}$ по web-сайту, расположенному на сервере.

3 Алгоритм формирования графа структуры сайта

Исходя из анализа структуры сайта, можно определить общий алгоритм формирования графа его структуры. Укрупненный алгоритм формирования графа структуры сайта представлен на рис. 3.

Алгоритм начинает работу с формирования главной страницы сайта A_1. Выполняется посылка запроса на главную страницу сайта (результатом запроса является Html-документ). По полученной информации формируется A_1. Также извлекается информация о содержимом страницы: заголовок, элементы управления, рисунки и др. (которые могут быть сохранены как свойства страницы), а также гиперссылки на другие страницы сайта a_{ij} (в случае первой итерации цикла разбора – a_{1j}).

Поиск гиперссылок производится в цикле. Каждая найденная ссылка a_{ij} проверяется на принадлежность к рассматриваемому серверу. Если данное условие выполняется, то производится проверка, не создана ли уже страница A_j (результирующая страница). Если нет, то A_j создается и помещается в очередь Q для более поздней обработки. Если же A_j существует, то у текущей рассматриваемой страницы добавляется ссылка a_{ij} и поиск продолжается.

После того, как все ссылки a_{ij} страницы A_i найдены, из очереди Q извлекается новый адрес A_j, снова посылается запрос и производится очередная итерация цикла алгоритма.

Алгоритм рекурсивен, для расчетов он использует очередь Q, в которой хранит адреса найденных страниц. Алгоритм заканчивает свою работу, когда очередь Q пуста.

В результате работы алгоритма формируется объект, представляющий модель сайта, с которой в дальнейшем можно осуществлять различные операции, такие, как сравнение, сортировка, поиск и др.

Реализация описанной методики требует описания предметной области в виде структуры реляционной базы данных.

Исходя из описанного алгоритма формирования графа структуры сайта, опишем средства, с помощью которых следует описать полученный объект.

Одним из наиболее используемых способов описания знаний является представление в виде фреймовой модели [5]. Его распространенность обусловлена близостью к концепции повсеместно применяемого объектно-ориентированного программирования. Представим структуру каждой вершины графа (структуры сайта) в виде фреймовой модели, изображённой на рис. 4.

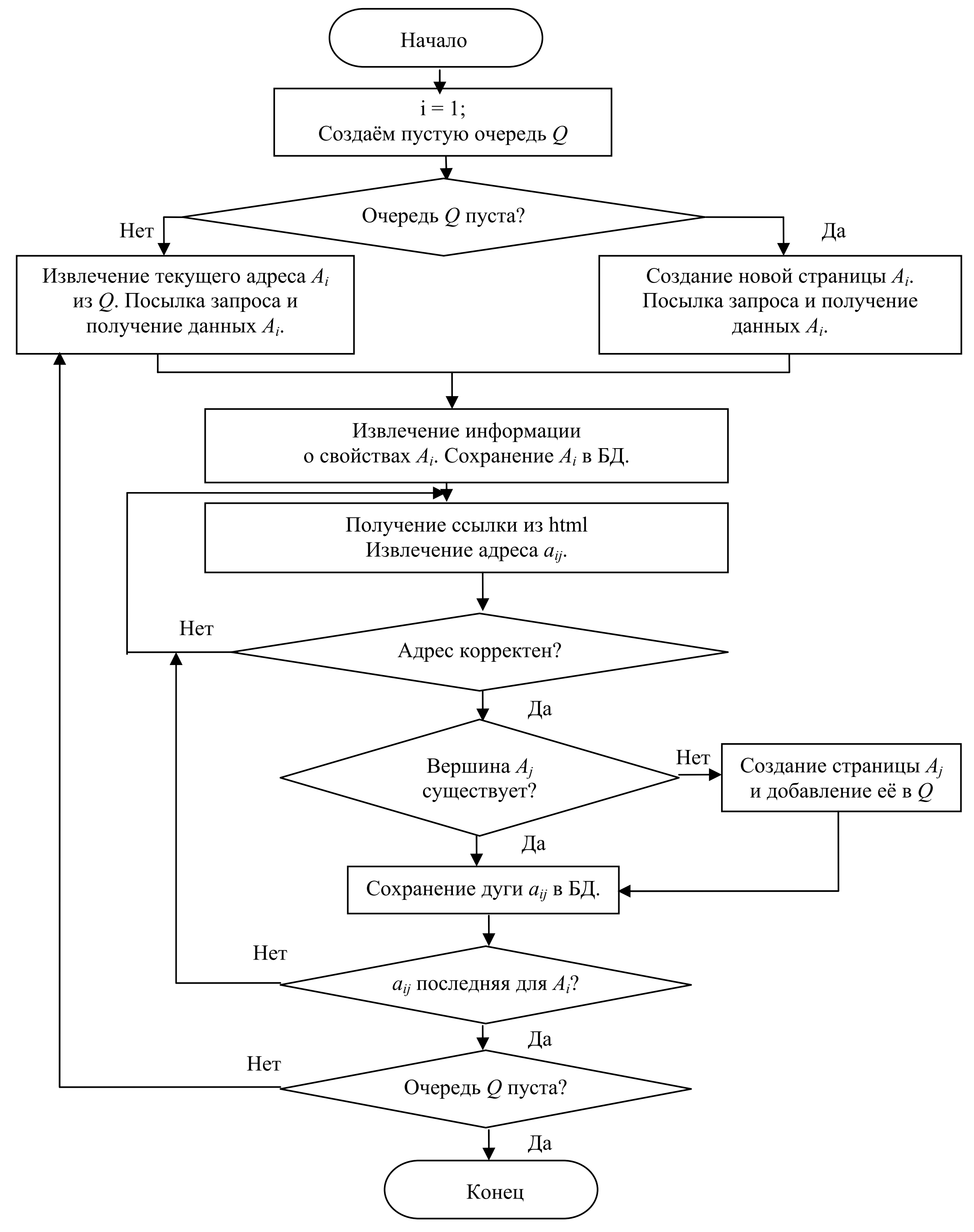


Рисунок 3 – Алгоритм формирования графа структуры сайта

Каждая вершина графа структуры сайта представляется в виде объекта, хранящего информацию о странице сайта A_i, наборе свойств этой страницы (например, элементов управления или других, в зависимости от поставленных задач) и дуг a_{ij}, принадлежащих A_i. В свою очередь, дуга a_{ij} – это объект, задаваемый индексами i и j ($i = 1, 2\ldots n$, $j = 1, 2\ldots n$) и значением ссылки (т.е. строковым значением адреса).

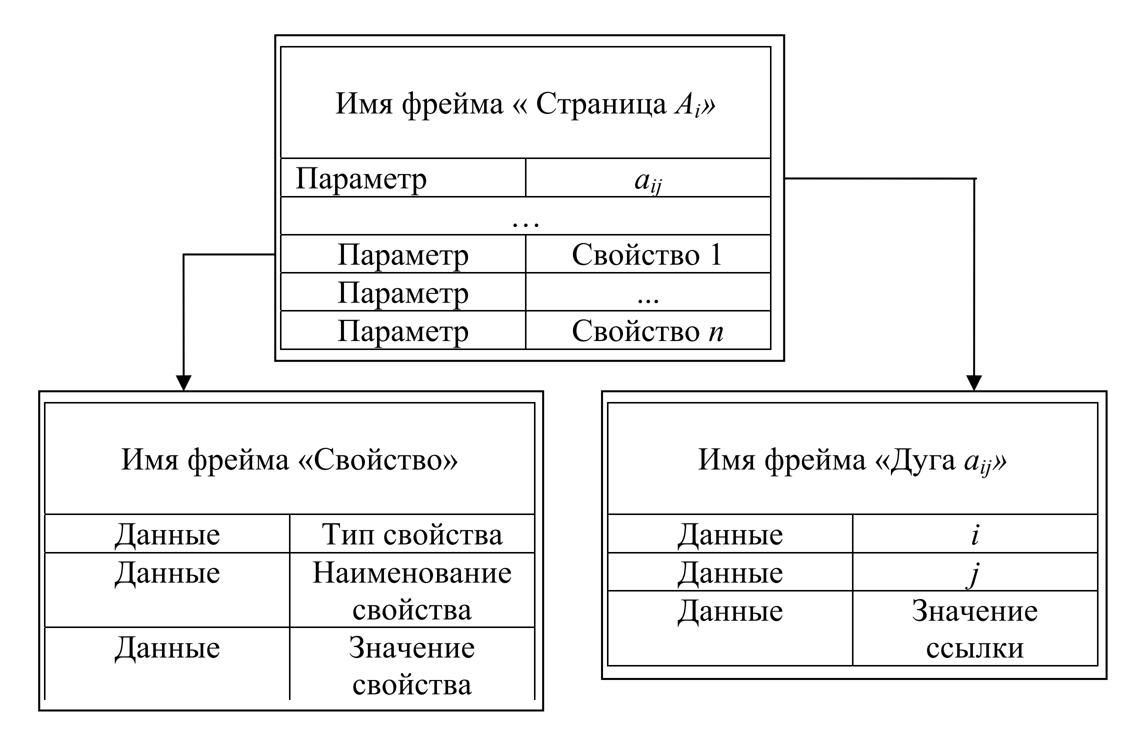


Рисунок 4 – Фреймовая модель для хранения информации о графе сайта

Свойства вершины A_i могут быть заданы различными способами в зависимости от того, какие именно свойства страницы необходимо обрабатывать в рамках конкретной задачи. В общем случае для каждого свойства следует задать тип, наименование и значение.

После разбора структуры сайта и представления его в виде графа, сайт может быть сохранён в реляционной базе данных для дальнейшей обработки [6]. Таким образом, появится возможность сохранять в базе данных множество сайтов для дальнейшего анализа, поиска, сравнения, выявления однотипных страниц и ссылочных структур и прочих операций.

В случае, если в базе данных сохранятся все свойства каждой web-страницы A_i, сайт впоследствии может быть целиком восстановлен в том виде, в каком он был сохранён в базе данных.

Выводы

В данной статье были рассмотрены принципы организации Интернет-сайтов и поставлена задача разработки эффективной модели представления структуры сайта для организации работы с ним, такой, как поиск по сайту, сравнение, сортировка или оценка сложности web-сайта.

Была рассмотрена структура web-сайтов и предложено несколько возможных моделей их представления – такие, как списковая модель и представление в виде графа. Было формализовано представление модели сайта в виде графа и разработан алгоритм формирования графа по сайту. Для реализации алгоритма были выделены структурные элементы сайта – страница и ссылка.

Для выполнения операций с web-сайтом из Интернет-клиентских программ, а также для сохранения сайта в реляционной базе данных (для дальнейшей групповой

обработки сайтов) приведено представление сайта в виде фреймовой модели. Предложенная модель является структурно расширяемой.

Данная методика может быть применена для написания программных модулей в Интернет-клиентских программах. Применение методики формирования объекта web-сайта позволит производить ряд операций над информацией, хранимой в Интернете.

Литература

1. Кристиансен Т., Торкингтон Н. Perl: Библиотека программиста: Пер. с англ. – СПб.: Питер, 2000. – 736 с.
2. Холзнер Стивен. Perl: специальный справочник: Пер. с англ. – СПб.: Питер, 2000. – 496 с.
3. Джейсон Мейнджер. Java: основы программирования: Пер. с англ. – К.: Издательская группа BHV, 1997. – 320 с.
4. Симкин Стив, Бартлет Нейл, Лесли Алекс. Программирование на Java. Путеводитель: Пер. с англ. – К.: НИПФ «ДиаСофт Лтд», 1996. – 736 с.
5. Эферган М. Java: справочник. – СПб.: Питер, 1998. – 448 с.
6. Хейл, Бернард Ван. JDBC: Java и базы данных: Пер. с англ. – М., 1999. – 320 с.

А.І. Ольшевський, А.А. Кондратьєва

Опис способів представлення web-сайтів у вигляді фреймової моделі для реалізації функціональних операцій в Інтернет-клієнтських системах

Розглянуто питання структурної побудови web-сайтів. Для реалізації алгоритму формування графа структури сайта було виділено елементи, запропоновано моделі, які дозволяють зберігати інформацію у реляційній базі даних для подальшої групової обробки в Інтернет-клієнтських програмах.

Статья поступила в редакцию 23.11.2007.